doi:10.1093/nar/gkp263

WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis

Daniel Glez-Peña[1], Gonzalo Gómez-López[2], David G. Pisano[2] and Florentino Fdez-Riverola[1,3,*]

[1]Higher Technical School of Computer Engineering, University of Vigo, Ourense, [2]Bioinformatics Unit (UBio), Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid and [3]Informatics Department, University of Vigo, Vigo, Pontevedra, Spain

Received January 30, 2009; Revised April 3, 2009; Accepted April 8, 2009

ABSTRACT

WhichGenes is a web-based interactive gene set building tool offering a very simple interface to extract always-updated gene lists from multiple databases and unstructured biological data sources. While the user can specify new gene sets of interest by following a simple four-step wizard, the tool is able to run several queries in parallel. Every time a new set is generated, it is automatically added to the private gene-set cart and the user is notified by an e-mail containing a direct link to the new set stored in the server. WhichGenes provides functionalities to edit, delete and rename existing sets as well as the capability of generating new ones by combining previous existing sets (intersection, union and difference operators). The user can export his sets configuring the output format and selecting among multiple gene identifiers. In addition to the user-friendly environment, WhichGenes allows programmers to access its functionalities in a programmatic way through a Representational State Transfer web service. WhichGenes front-end is freely available at http://www.whichgenes.org/, WhichGenes API is accessible at http://www.whichgenes.org/api/.

INTRODUCTION

During the past several years, bioinformatics enrichment tools have played a very important and successful role contributing to the gene functional analysis of large gene lists (ranging in size from hundreds to thousands of genes) for various high-throughput biological studies (1).

From the large amount of tools that are currently available in the community, two widely used approaches can be identified: (i) individual gene analysis (IGA), which evaluates the significance of individual genes between two groups of samples compared, and (ii) gene set analysis (GSA), free from the problems of the 'cutoff-based' methods (2).

In recent years, GSA approach has received a great deal of attention because, from a biological perspective, functionally related genes often display a coordinated expression to accomplish their roles in the cell. In this direction, GSA methods enable the understanding of cellular processes as an intricate network of functionally related components (3). However, while extensive work has been done during last years in developing new GSA methods, little effort has been put on implementing tools that can help researches gather, store and manage gene sets containing large 'interesting' gene lists from multiple data sources. To our knowledge, only the Gene Set Builder tool (4) shares the same fundamental concept, giving support to easily handle sets of genes. Compared with Gene Set Builder, which follows a fixed database-driven architecture to give access only to Ensembl and GeneLynx gene catalogs, the tool presented herein follows a more flexible database-free and interactive approach allowing the integration of 14 different data sources for the *Homo sapiens* and *Mus musculus* organisms.

In this article, we present WhichGenes, an on-line, database-free, web-based tool for easily gathering, building, storing and exporting always updated gene sets coming from multiple data sources. It allows researchers to elaborate custom hypotheses in the form of lists of genes in order to further use them as input in existing GSA tools. WhichGenes currently supports queries about *Homo sapiens* and *Mus musculus* organisms by retrieving up-to-date gene lists directly coming from multiple databases, currently including Ensembl, MSigDB, KEGG/Biocarta/Reactome pathway databases, GeneCards, CancerGenes, Decipher, Diseases CTD, TargetScan, miRBase, Chemical CTD, AmiGO and IntAct. Generated gene sets can be

*To whom correspondence should be addressed. Tel: +34 988 387015; Fax: +34 988 387001; Email: riverola@uvigo.es easily combined and exported using a wide variety of supported gene identifiers. In addition, WhichGenes provides a web service allowing the server to be directly accessed from multiple programming languages.

MATERIALS AND METHODS

Gene set data sources

In GSA as important as the algorithms are the gene sets. WhichGenes retrieves and integrates different data for constructing target gene sets by accessing diverse sources of biological knowledge. Each gene set is created running a given query over the user specified data source. There are two types of queries depending on the source of information: (i) free-text query data sources and (ii) catalogue-based query data sources. In the first case, the user specifies a query using a text box, i.e. writing 'leukemia' in order to retrieve those genes related to this disease from 'GeneCards'. In the second case, catalogue-based queries force the user to select one or more terms from a fixed catalogue displayed in a list or a tree. Catalogue-based queries are used, for example, to retrieve genes annotated with a particular GO term that are involved in a desired pathway or found in a specific chromosome location. Currently, WhichGenes gives access to 14 different data sources of two species (*Homo sapiens* and *Mus musculus*) as illustrated in Table 1.

System workflow

The WhichGenes usage workflow typically requires iteration between two simple stages: (i) gene sets creation phase and (ii) gene-set management phase. Figure 1 shows a summary of the main procedures carried out in each stage.

The process to create gene sets starts with the selection of the desired data source (categorized by organism). If the selected data source supports catalogue-based queries (i.e. pathway data sources), the closed set of terms (pathway names) are retrieved and displayed to the user. Since the catalogue depends on the data source and can be dynamically updated, its terms are retrieved on-demand for each query. In the next step, the user configures a query by typing text in case of free-text queries, or by selecting one or more terms in case of catalogue-based queries. Finally, the user is asked to specify a name for the new set. After these steps, WhichGenes runs the query and on-line retrieves a list of related genes from the data source. Following this simple procedure in an iterative way, the user can perform multiple queries simultaneously with the goal of creating several sets. The genes coming from each data source are available via different identifiers (i.e. KEGG provides Entrez IDs, AmiGO uses Uniprot IDs, miRBase returns Ensembl IDs, etc.). In order to standardize all the sets and use the same namespace, WhichGenes accesses the Ensembl BioMart MartService to query the *Homo sapiens* and *Mus musculus* gene datasets in order to convert previous retrieved identifiers to HGNC and MGI symbols (for human and mouse organisms, respectively).

Every time WhichGenes standardizes the identifiers corresponding to a recent generated list of genes, a new set is automatically added to the private 'gene-set cart' and an e-mail is sent to the user with a direct link for easily accessing and sharing the set. In the gene-set management phase, WhichGenes provides to the user several operations for manipulating existing sets. These actions facilitate both (i) the modification of individual sets by deleting particular genes or deleting/duplicating/renaming the whole set and (ii) the combination of existing sets for generating new synthesized ones. The last group of actions has special relevance since it allows the user to merge two or more sets by using the implemented '*and*', '*or*' and

Table 1. Data sources used by WhichGenes to automatically construct gene sets

Source description	Species[a]	Catalogue-based	Access method
Positional: genes located in user specified chromosomic ranges			
Ensembl (5)	H, M	Yes—list	BioMart API
MSigDB Positional GeneSets (6)	H	Yes	Web parsing
Pathways: genes involved in user specified pathways			
KEGG (7)	H, M	Yes—list of pathways	KEGG web service API
Biocarta (8)	H, M	Yes—list of pathways	Local database
Reactome (9)	H, M	Yes—list of root pathways	BioMart API
Diseases: genes related to user specified diseases			
GeneCards (10)	H	No	Web parsing
Cancer Genes (11)	H	Yes—list of projects	Web parsing
Decipher (12)	H	Yes—list of syndromes	Web parsing
CTD (13)	H, M	No	Web parsing
MicroRNA targets: genes that are target of user specified microRNAs			
TargetScan (14)	H	No	Web parsing
miRBase (15)	H, M	No	Web parsing
Chemical: genes interacting with user specified chemicals			
CTD (13)	H, M	No	Web parsing
Other:			
AmiGO (16): genes annotated with user specified GO terms	H	Yes—tree of GO terms	Web parsing
IntAct (17): genes interacting with the user specified genes	H	No	Web parsing

[a]'H' stands for *Homo sapiens* and 'M' for *Mus musculus* organisms. WhichGenes supports different access methods using standard APIs where possible or applying web parsing where information is not programmatically accessible.

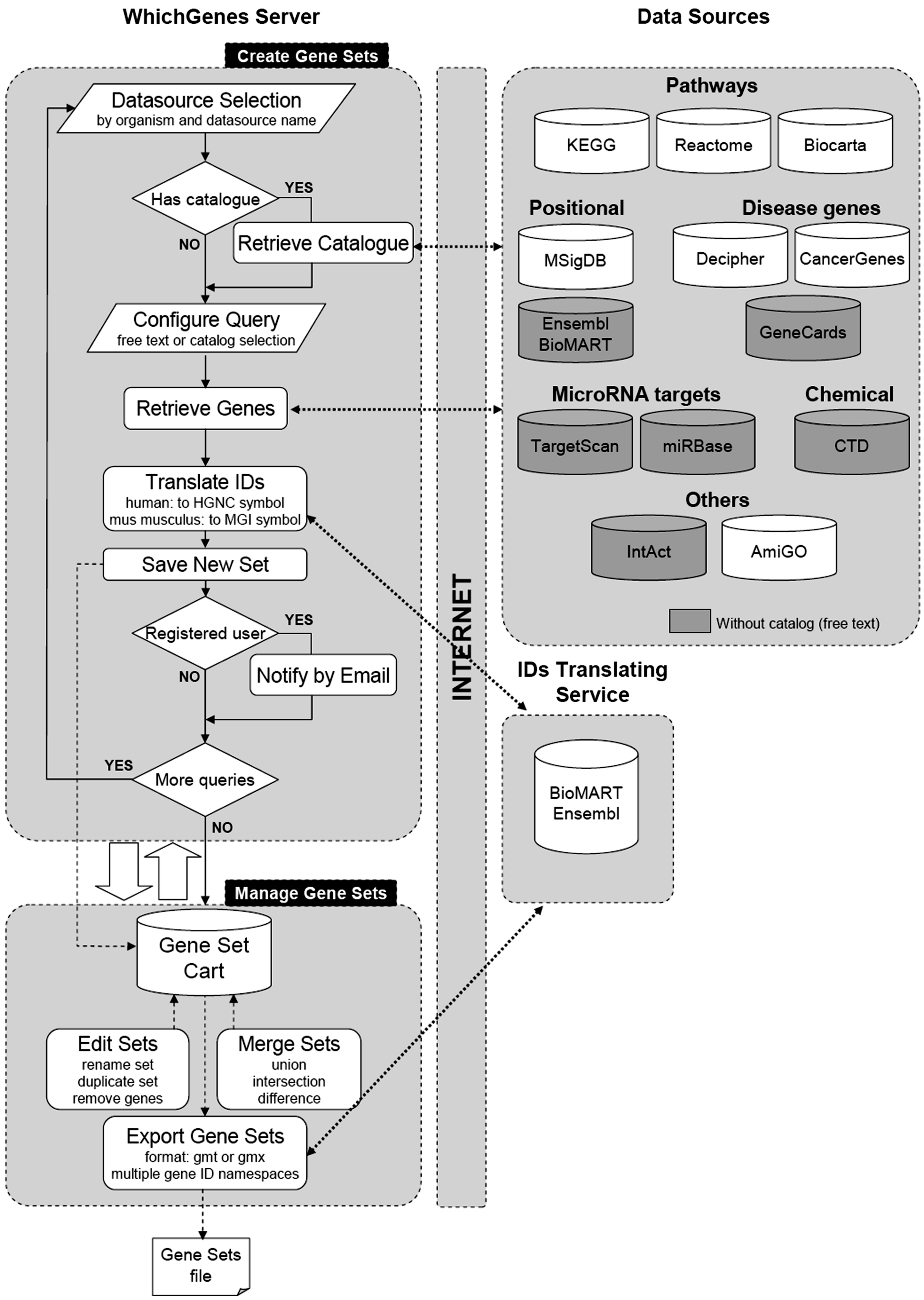


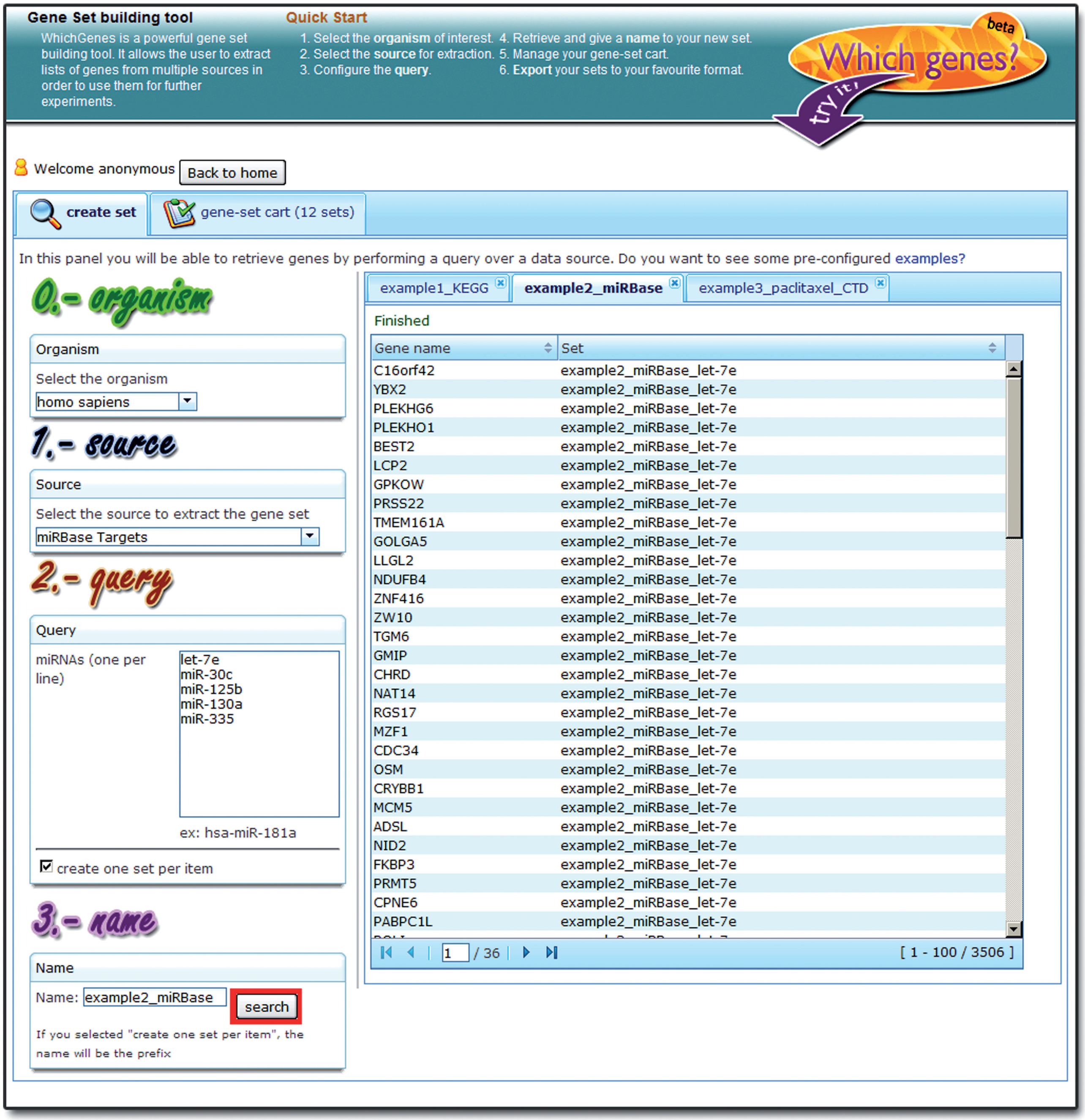
Figure 1. Schematic WhichGenes workflow. The user can iterate between gene set creation and gene set management phases before exporting the desired group of gene sets.

'*difference*' operators. This feature makes possible to create new sets supporting more elaborated hypotheses (i.e. genes belonging to a given pathway AND targeted by a particular miRNA). Through the logical operations of gene sets between different functional classifications, interesting gene lists can be separated into more specific and smaller groups of genes, which further facilitate a much more detailed analysis of expression patterns.

Once the users obtain the desired gene sets supporting their custom hypotheses, they can be exported by applying two commonly used text-based formats (CSV and GSEA). It is also possible to change the gene namespace employed in the exported files by using the implemented IDs Translating Service.

Implementation

WhichGenes is implemented as an AJAX-enabled web application programmed in J2SE 1.5 Java language. The ZK development framework (http://www.zkoss.org) was used to construct the user interface as well as different technologies to access the source databases including the



Gene Set building tool

WhichGenes is a powerful gene set building tool. It allows the user to extract lists of genes from multiple sources in order to use them for further experiments.

Quick Start

1. Select the **organism** of interest.
2. Select the **source** for extraction.
3. Configure the **query**.
4. Retrieve and give a **name** to your new set.
5. Manage your gene-set cart.
6. **Export** your sets to your favourite format.

Welcome anonymous [Back to home]

create set | gene-set cart (12 sets)

In this panel you will be able to retrieve genes by performing a query over a data source. Do you want to see some pre-configured examples?

0.- organism

Organism

Select the organism: homo sapiens

1.- source

Source

Select the source to extract the gene set: miRBase Targets

2.- query

Query

miRNAs (one per line)

```
let-7e
miR-30c
miR-125b
miR-130a
miR-335
```

ex: hsa-miR-181a

☑ create one set per item

3.- name

Name

Name: example2_miRBase [search]

If you selected "create one set per item", the name will be the prefix

example1_KEGG | **example2_miRBase** | example3_paclitaxel_CTD

Finished

Gene name	Set
C16orf42	example2_miRBase_let-7e
YBX2	example2_miRBase_let-7e
PLEKHG6	example2_miRBase_let-7e
PLEKHO1	example2_miRBase_let-7e
BEST2	example2_miRBase_let-7e
LCP2	example2_miRBase_let-7e
GPKOW	example2_miRBase_let-7e
PRSS22	example2_miRBase_let-7e
TMEM161A	example2_miRBase_let-7e
GOLGA5	example2_miRBase_let-7e
LLGL2	example2_miRBase_let-7e
NDUFB4	example2_miRBase_let-7e
ZNF416	example2_miRBase_let-7e
ZW10	example2_miRBase_let-7e
TGM6	example2_miRBase_let-7e
GMIP	example2_miRBase_let-7e
CHRD	example2_miRBase_let-7e
NAT14	example2_miRBase_let-7e
RGS17	example2_miRBase_let-7e
MZF1	example2_miRBase_let-7e
CDC34	example2_miRBase_let-7e
OSM	example2_miRBase_let-7e
CRYBB1	example2_miRBase_let-7e
MCM5	example2_miRBase_let-7e
ADSL	example2_miRBase_let-7e
NID2	example2_miRBase_let-7e
FKBP3	example2_miRBase_let-7e
PRMT5	example2_miRBase_let-7e
CPNE6	example2_miRBase_let-7e
PABPC1L	example2_miRBase_let-7e

1 / 36 [1 - 100 / 3506]

Figure 2. WhichGenes 'create set' wizard running a user query. Left panel shows the specified parameters. Right panel visualizes the retrieved genes as soon as they are available using different tabs for each query. When the list of retrieved genes exceeds a predefined threshold, WhichGenes automatically paginates the results.

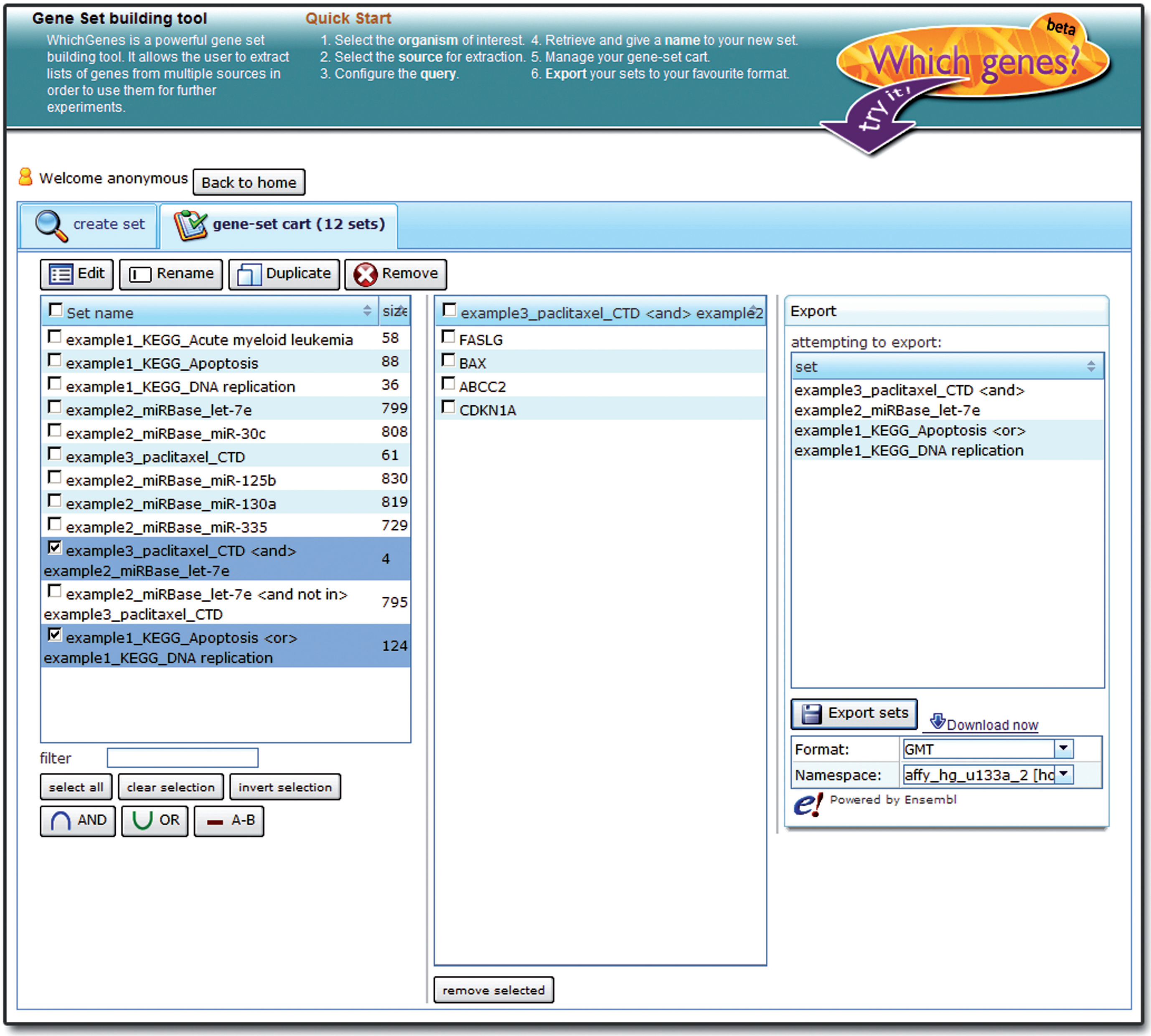
BioMart XML-based query service for Ensembl and Reactome, the KEGG API and a custom developed library to download, parse, process and extract information from unstructured HTML pages, since many databases can only be accessed via their web sites. We have also implemented caching techniques in order to reduce the number of accesses to the external data sources, especially when retrieving catalogues.

In addition to the functionality provided by the end user front-end, WhichGenes also implements a REST-based programming interface allowing developers to retrieve gene sets directly from our server and using them within their own algorithms.

WhichGenes runs on a standard Tomcat 5.5 Web application server without any local database. An anonymous login is available for those who are interested in testing the system (without e-mail notifications) as well as preloaded searches.

Supported platforms

Currently, WhichGenes has been successfully tested in Internet Explorer 7, Firefox 3, Opera 9.62 and Safari 3



Gene Set building tool

WhichGenes is a powerful gene set building tool. It allows the user to extract lists of genes from multiple sources in order to use them for further experiments.

Quick Start

1. Select the **organism** of interest.
2. Select the **source** for extraction.
3. Configure the **query**.
4. Retrieve and give a **name** to your new set.
5. Manage your gene-set cart.
6. **Export** your sets to your favourite format.

□ Set name	size
□ example1_KEGG_Acute myeloid leukemia	58
□ example1_KEGG_Apoptosis	88
□ example1_KEGG_DNA replication	36
□ example2_miRBase_let-7e	799
□ example2_miRBase_miR-30c	808
□ example3_paclitaxel_CTD	61
□ example2_miRBase_miR-125b	830
□ example2_miRBase_miR-130a	819
□ example2_miRBase_miR-335	729
☑ example3_paclitaxel_CTD <and> example2_miRBase_let-7e	4
□ example2_miRBase_let-7e <and not in> example3_paclitaxel_CTD	795
☑ example1_KEGG_Apoptosis <or> example1_KEGG_DNA replication	124

□ example3_paclitaxel_CTD <and> example2
□ FASLG
□ BAX
□ ABCC2
□ CDKN1A

Export

attempting to export:

set
example3_paclitaxel_CTD <and> example2_miRBase_let-7e
example1_KEGG_Apoptosis <or> example1_KEGG_DNA replication

Format:	GMT
Namespace:	affy_hg_u133a_2 [hc

Figure 3. WhichGenes 'gene-set cart'. Left panel shows the gene set list where gene sets can be removed, renamed and merged. Middle panel is used to enumerate the genes belonging to the selected set and supports filtering and deleting operations. Right panel provides export functionalities where the user specifies the desired output format and gene IDs namespace.

browsers working on Windows XP/Vista, Ubuntu Linux 8.04 version and Mac OSX 10.5 of Intel architecture.

RESULTS AND DISCUSSION

WhichGenes presents a natural web interface based on a simple to use four-step wizard to create groups of interesting genes plus an intuitive control panel to manage and export them. The 'create set' wizard allows the user to select the organism, specify the data source, configure the query and finally, give a name to the set. An example of a search is provided in Figure 2. When the search starts, its progress can be observed in a panel on the right zone. While the search is generating the results the user can configure and launch new queries, which will run simultaneously.

Every time a new set is generated by WhichGenes, it is automatically added to the private 'gene-set cart' where the user can manage his own groups. Figure 3 shows the functionalities of this control panel.

In order to always manage updated gene sets, WhichGenes follows a database-free (or federated) architecture. This approach has several advantages including the fact of always retrieving up-to-date information and the absence of periodically duplicating huge databases. In addition, the required source data and/or query algorithms could be not always fully accessible to download in order to replicate them in our server. However, our approach also poses some disadvantages related with a poorer performance, the possibility of temporal unavailable third-party databases, and the need to update the search agents that parse and reformat results whenever a source database change their data structure (18). Nevertheless, in our daily usage we have verified a similar performance of retrieving genes from WhichGenes in comparison to the original data sources. We also check for the availability of external databases by running a sample query over every data source in a daily basis, disabling it and notifying the administrator whenever it fails. Finally, our search agents are designed with an in-house specific web crawler/scraping library, which facilitates both the creation and modification of their code.

WhichGenes provides a set of unique features which are not currently available in other tools: a very intuitive interface, up-to-date access to supported databases, many sources of information, possibility of downloading several sets of genes in an unique file, support for standard output files for GSEA (.gmt,.gmx), possibility of exploring subtrees from certain sources (i.e. GO) in order to make more specific queries, etc.

CONCLUSIONS

The comprehensive functional analysis of large gene lists, derived in most cases from emerging high-throughput genomic, proteomic and bioinformatics scanning approaches, is still a challenging and daunting task. While a significant number of bioinformatics enrichment tools are currently available in the community, few efforts has been invested in developing tools for gathering, building, storing and managing gene sets containing large 'interesting' gene lists. WhichGenes is specifically focussed in this area by offering final-user facilities as well as a programmatic API for intuitively extracting lists of genes from multiple sources in order to extent the scope of existing gene set based analysis methods.

WhichGenes is able to access and integrate information coming from a continuously growing plethora of information repositories including pathways, diseases, chemicals, GO, microRNAs, etc. The 'database-free' nature of WhichGenes implies that external data sources are accessed just in time, retrieving up-to-date information without using obsolete local mirrors of existing data sources. The user can not only retrieve and integrate 'interesting' gene lists from multiple repositories automatically, but also combine these sets (with the '*and*', '*or*' and '*difference*' operators) to build more complex hypotheses. As different namespaces are commonly used, WhichGenes automatically standardizes all the generated sets to HGNC and MGI symbols in order to correctly perform merging operations. However, WhichGenes can export gene sets in commonly used text-based formats where the final gene namespace can be changed to a more appropriate one, including multiple popular database gene identifiers and several microarray probe set IDs.

FUNDING

This work was initiated and partially funded by the project *Research on Translational Bioinformatics* (08VIB6) from University of Vigo. María Barbeito contract from Xunta de Galicia (to D.G.-P.); Spanish National Institute for Bioinformatics, a platform of Genoma España (to G.-G.-L.). Funding for open access charge: Informatics Department, University of Vigo.

Conflict of interest statement. None declared.

REFERENCES

1. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2008) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
2. Nam,D. and Kim,S-Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinformatics*, **9**, 189–197.
3. Dopazo,J. (2006) Functional interpretation of microarray experiments. *OMICS*, **10**, 398–410.
4. Yusuf,D., Lim,J.S. and Wasserman,W.W. (2008) The Gene Set Builder: collation, curation, and distribution of sets of genes. *BMC Bioinformatics*, **6**, 305.
5. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
6. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Elbert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550.
7. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
8. Biocarta. (2008) Charting Pathways of Life. http://www.biocarta.com/. (last accessed on January 29, 2009)
9. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase for human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
10. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,E., Feldmesser,G. *et al.* (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31(Database issue)**, 142–146.
11. Higgins,M.E., Claremont,M., Major,J.E., Sander,C. and Lash,A.E. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.*, **35**, D721–D726.
12. Decipher: DatabasE of Chromosmoal Imbalance and Phenotype in Humans using Ensembl Resources. (2008) http://decipher.sanger.ac.uk/. (last accessed on January 29, 2009)
13. Mattingly,C.J., Rosenstein,M.C., Davis,A.P., Colby,G.T., Forrest,J.N. and Boyer,J.L. (2006) The comparative Toxicogenomics Database: a cross-species resource for building chemical–gene interaction networks. *Toxicol Sci.*, **92**, 587–595.
14. Benjamin,P.L., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
15. Griffiths-Jones,S., Saini,H.K., Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
16. The Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
17. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
18. Zhang,Z., Cheung,K. and Townsend,J.P. (2009) Bringing Web 2.0 to bioinformatics. *Brief. Bioinformatics*, **10**, 1–10.